PB89-134712

Target Transformation Factor Analysis as an Aerosol Mass Apportionment Method A Review and Sensitivity Study (Journal Version)

Illinois Univ. at Urbana-Champaign

Prepared for

Environmental Protection Agency
Research Triangle Park, NC

1988

U.S. Department of Commerce
National Technical Information Service
NTIS

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before complet

1. REPORT NO. EPA/600/J-88/213	2.
3. PB89-134712	
4. TITLE AND SUBTITLE TARGET TRANSFORMATION FACTOR ANALYSIS AS AN AEROSOL MASS APPORTIONMENT METHOD: A REVIEW AND SENSITIVITY STUDY (Journal Version)	5. REPORT DATE 1988
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Philip K. Hopke	8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Institute for Environmental Studies University of Illinois 1005 W. Western Avenue Urbana, Illinois 61801	10. PROGRAM ELEMENT NO. CDTA1D/01-0357
	11. CONTRACT/GRANT NO. 68-02-3449
12. SPONSORING AGENCY NAME AND ADDRESS Atmospheric Sciences Research Laboratory - RTP, NC Office of Research and Development U.S. Environmental Protection Agency Research Triangle Park, North Carolina 27711	13. TYPE OF REPORT AND PERIOD COVERED
	14. SPONSORING AGENCY CODE EPA/600/09

15. SUPPLEMENTARY NOTES
REF: Atmospheric Environment, 22(9):1777-1792, 1988

16. ABSTRACT

Over the past several years, a multivariate receptor model called Target Transformation Factor Analysis (TTFA) has been applied to various problems of environmental source resolution. In recent reports on the analysis of data sets with known structure, good results were obtained. In order to further test the TTFA approach, additional simulated data sets have been created with varying degrees of collinearity of the source profiles and normally distributed uncertainty. The effect of the collinearity is to restrict the number of resolvable sources. The analytical error affects the identification of the number of sources and development of accurate profiles.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group

REPRODUCED BY
U.S. DEPARTMENT OF COMMERCE
NATIONAL TECHNICAL INFORMATION SERVICE
SPRINGFIELD, VA. 22161

18. DISTRIBUTION STATEMENT RELEASE TO PUBLIC	19. SECURITY CLASS (This Report) UNCLASSIFIED	21. NO. OF PAGES 18
	20. SECURITY CLASS (This page) UNCLASSIFIED	22. PRICE

EPA/600/J-88/213
JOURNAL ARTICLE

Atmospheric Environment Vol. 22, No. 9, pp. 1777–1792, 1988.
Printed in Great Britain.

0004-6981/88 $3.00 + 0.00
© 1988 Pergamon Press plc

NTIS is authorized to reproduce and sell this report. Permission for further reproduction must be obtained from the copyright owner.

TARGET TRANSFORMATION FACTOR ANALYSIS AS AN AEROSOL MASS APPORTIONMENT METHOD: A REVIEW AND SENSITIVITY STUDY

PHILIP K. HOPKE

*Institute for Environmental Studies, University of Illinois, 1005 W. Western Avenue, Urbana, IL 61801, U.S.A. and Farmaceutisch Instituut, Vrije Universiteit Brussel, Laarbeeklaan 103, B-1090 Brussels, Belgium

(*First received* 25 *April* 1985 *and received for publication* 7 *March* 1988)

Abstract—Over the past several years, a multivariate receptor model called Target Transformation Factor Analysis (TTFA) has been applied to various problems of environmental source resolution. In recent reports on the analysis of data sets with known structure, good results were obtained. In order to further test the TTFA approach, additional simulated data sets have been created with varying degrees of collinearity of the source profiles and normally distributed uncertainty. The effect of the collinearity is to restrict the number of resolvable sources. The analytical error affects the identification of the number of sources and development of accurate profiles.

Key word index: Target transformation, spectra analysis, source apportionment, sensitivity allowances.

INTRODUCTION

In 1971, the first National Ambient Air Quality Standards were established, setting limits on permissible levels of total suspended particulate matter. Because these regulations concentrated only on the absolute amounts of suspended particulate matter, control efforts were directed primarily toward the removal of the larger, more massive particles that comprise most of the total mass of particle emissions. There is growing evidence, however, that the smaller particles, although representing a smaller fraction of the total mass, present the greater threat to public health and welfare. Because of their light scattering properties, the smaller particles also contribute to a larger extent to the reduction in visibility. Concern over the possible health effects of smaller-diameter particulate matter is reflected in the development and use of aerosol sampling equipment that measures the concentrations of particles in the inhalable (diameters $< 10\ \mu m$) and fine (diameters $< 2\ \mu m$) particle-size fractions. New regulations, recently promulgated by the EPA, include standards on the levels of inhalable particles. The effectiveness of regulations designed to limit the impact of fine and inhalable particles on the public would be greatly enhanced if the sources and relative emission rates of these particles were well known.

BACKGROUND

One approach to the problem of determining the contribution of each source to the total level of suspended particulate matter is the method of source-emission rates for all known sources along with a dispersion model for their behavior following emission. This method attempts to determine the contribution of each source to the total aerosol concentration. The results of source-emissions inventories are too often unsatisfactory due to a lack of accurate source-emission profiles and the inability to include fugative sources. The development of improved aerosol sampling equipment has permitted the development of receptor methods for the determination of source contributions. These methods employ statistical analysis techniques to calculate the contribution of each source of particulate matter from a representative set of ambient aerosol samples.

The fundamental assumption of receptor models is that the measured concentration of a particular species is the result of a linear sum of independent contributions by each source, natural or anthropogenic, that produces the particular measured chemical species. For example, the total particulate Pb concentration in a given parcel of air can be considered as the sum of lead-containing particles from motor vehicles burning leaded fuel, from refuse incinerators, from non-ferrous metal smelters, etc. Since particles from a given source generally contain more than one species, the airborne concentration is the product of two cofactors; the concentration of the species in the material emitted by the source and the total mass loading contributed by the source to the particular air parcel sampled. For a number of species, the mass balance can then be summarized by

$$x_{ij} = \sum_{k=1}^{p} a_{ik} f_{kj} \qquad (1)$$

where x_{ij} is the volumetric concentration of the ith species identified in the jth sample, a_{ik} is the gravimetric concentration of the ith species in material from

* Permanent address.

the kth source, and f_{kj} is the volumetric mass contribution of the kth source to the jth sample. Thus, the receptor model is trying to express the measured ambient concentration in this finite linear expansion in the source profile vectors a_k.

In the first applications of these methods, it was assumed that the general source classes and their properties were known. Early applications of what was called a Chemical Element Balance (CEB) were to ambient aerosol samples from Pasadena, California (Miller *et al.*, 1972; Friedlander, 1973), Ghent, Belgium (Heindryckx and Dams, 1974), Heidelberg, F.R.G. (Bogen, 1973), and Chicago, Illinois (Gatz, 1975). These early applications of the chemical element balance method suffered from a lack of reliable data for the source emission profiles needed for the analysis. Improvements in the quality of source profile determinations (Greenberg, 1976; Ondov, 1974; Mroz, 1976; Gladney, 1974; Thomae, 1977) have produced much better agreement between the calculated and measured elemental concentrations for airborne particulate matter in Washington, D.C. (Kowalczyk *et al.*, 1982) and for dichotomous sampler data in St. Louis and the Great Smoky Mountains (Dzubay, 1980). The scope of the chemical element balance method was further expanded in an extensive study of the Portland, Oregon aerosol as part of the PACS program (Watson, 1979; Kowalczyk *et al.*, 1979; Watson *et al.*, 1984). This study included components such as vegetative and grass burning and chemical species such as volatile and nonvolatile C, NO_3^-, and SO_4^{2-} that had not previously been used. By including the larger number of source components, very good agreement was achieved for the 27 elements and chemical species that had been measured. Chemical element balance methods have a major drawback in that they require an *a priori* knowledge of both the number and composition of the suspected source emissions. The only quantities calculated are the contribution of each of the presumed sources to each of the measured samples. Though extensive efforts by the groups in Maryland and Oregon have produced much better determinations of the elemental profiles of a number of sources, these determinations are time consuming, expensive and, in many cases, inapplicable to the resolution of aerosol sources in different locations around the United States. In addition, source determinations of in-stack material cannot account for chemical and physical transformations of volatile or reactive species following emission.

An alternative method has been suggested where the necessary basis vectors for the expansion in Equation (1) are derived directly from the data. These eigenvector methods are generally refered to as principal components and/or factor analysis. These methods have come into use primarily from the statistical analysis of data. However, caution must be exercised in viewing these as 'statistical' methods for air quality data. In general, the purpose of a statistical analysis is to be able to infer the properties of a population from a set of samples selected from that population. There are constraints on the nature of these samples in that it is necessary that they be identical and independent samples taken from a multivariate normal distribution. Under these circumstances, one can statistically analyze the data and infer the properties of the entire population. With typical air quality data, these assumptions are clearly violated and, therefore, the analysis is being made to understand the nature of the particular samples that comprise the data set. Any extension of the interpretation beyond these data is not statistically justified.

Several early studies made use of the classical factor analysis methods that had been developed for statistical analysis purposes. Blifford and Meeker (1967) and Prinz and Stratmann (1968) were earlier efforts to interpret air pollution data with factor analysis. However, the quality of data available to them made the analysis difficult to interpret.

Hopke *et al.* (1976) used factor and cluster analyses to identify sources of airborne particulate matter in the Boston urban aerosol. In this study, the data were first transformed into a standardized form that removed the effects of using different metrics in describing the data. The identification of sources was made by associating the largest calculated factor loading with the marker elements for each source. In a similar fashion, factor and cluster analyses were used to characterize the aerosol of Tucson, Arizona (Gaarenstroom *et al.*, 1977). Gatz (1978) used factor analysis to identify sources of particulate matter in St. Louis. Hopke *et al.* (1976) remarked in their report that this form of factor analysis can identify the major causes of the variation in the elemental concentration but did not lead directly to calculation of the mass contributions as could be obtained from a mass balance analysis. Subsequently, Thurston and Spengler (1985) have developed a method to obtain the mass contributions. The principal components analysis yields a model that reproduces the standardized data, z_{ij},

$$z_{ij} = (x_{ij} - \bar{x}_i)/\sigma_i \tag{2}$$

where x_{ij} is the original data value, $\bar{x}_i$ is the mean value of that variable over the data set and σ_i is its standard deviation. The data have thus been centered around the mean value and the true zero in the system has been lost. Although their method can provide an actual mass apportionment, these additional steps lead to increased uncertainties in the resulting values caused by the bias in estimating the mean of nonsymmetrically distributed values (Curran and Suggs, 1986).

TARGET TRANSFORMATION FACTOR ANALYSIS

An analysis that would actually apportion the mass contributions can be obtained by utilizing an alternative approach to the development of the set of basis

vectors from the data. Several studies by Hopke and co-workers (Hopke *et al.*, 1980; Alpert and Hopke, 1980, 1981; Hopke, 1980; Liu *et al.*, 1982; Severin *et al.*, 1983) have employed an alternative form of factor analysis called target transformation factor analysis (TTFA). Originally developed by Weiner *et al.* (1970), this technique permits an analysis, similar in nature to the chemical element balance method by deriving source profile vectors that are not standardized. The approach has been described in seve al publications (Roscoe and Hopke, 1981a,b; Hopke *et al.*, 1983) and is presented by Hopke (1985). Therefore, only an outline of the procedure will be presented here.

The objective of TTFA is to resolve the ambient data set into source contributions given by the matrix equation

$$X = AF \tag{3}$$

where A is a matrix of dimension $m \times p$ containing the concentration profile vectors for each of the p sources and F, a $p \times n$ matrix, is composed of the amounts per unit volume that each source contributes to each sample. The goals of a factor analysis are: (1) to determine p, the number of independent sources that contribute to the system, (2) to determine the components of matrix A, the elemental source profiles, and (3) to calculate F, the contribution of each source to each sample. The questions arise as to how accurately can this procedure analyze ambient air quality data and what are the limitations that exist in this method.

Mathematical procedure

The first step in the analysis is the development of the basis set that is going to be used to represent the source contributions to the observed data. In order to remove the different scale of measurement for the various chemical species, the data is scaled to yield a correlation matrix. In several reports (Hwang *et al.*, 1984; Roscoe *et al.*, 1984) a distinction has been made as to the form of the correlation matrix that is used to extract the eigenvectors. If the data matrix, x, is m elements by n samples, then two correlation matrices can be obtained depending on the order of multiplication of the data matrix and its transpose. In what has been termed a Q-mode analysis, the correlation matrix is given by

$$(XV_Q^{-1})^t(XV_Q^{-1}) = C_Q \tag{4}$$

where V_Q is a diagonal matrix with elements given by

$$v_{jj} = \left[\sum_{i=1}^{m} x_{ij}^2\right]^{1/2}. \tag{5}$$

Alternatively,

$$(V_R^{-1}X)(V_R^{-1}X)^t = C_R$$

$$v_{ii} = \left[\sum_{j=1}^{n} x_{ij}^2\right]^{1/2}. \tag{6}$$

However, it has recently been pointed out (Heidam and Kronberg, 1985) that the reason that different results are obtained from the analysis of C_Q and C_R is not the order of the matrix multiplication but rather the direction of the scaling. The effect of the type of data scaling on the factor analysis solution has also been discussed by Zhou *et al.* (1983). They also show that the R- and Q-mode solutions can both be interrelated through a singular value decomposition of the data matrix.

$$X = V_R SDT^t = SDT^t V_Q \tag{7}$$

where S is the $m \times m$ matrix of eigenvectors obtained from the XX^t matrix, T is the $n \times n$ matrix of eigenvectors obtained from the X^tX matrix and D is the $m \times n$ diagonal matrix containing the square roots of the first n eigenvectors obtained from the X^tX matrix. Thus, the values of the two modes are directly interrelated and the values of one mode can be calculated from the other. There still are differences in resulting A matrices largely because of the multiplication of the scaling factors. Thus, the R- and Q-mode designation based on the dimension, $m \times m$ or $n \times n$, of the correlation matrix is not the critical distinction and it is direction of scaling that is the point of difference.

The correlation matrix is next diagonalized by finding a matrix such that

$$Q^{-1}CQ = \Lambda \tag{8}$$

where Λ is a diagonal matrix of eigenvalues arranged in decreasing order of magnitude and Q contains the associated eigenvectors. Depending on whether the R- or Q-mode correlation matrix is analyzed, the values of the A and F matrices can be calculated using the equations given by Hwang *et al.* (1984).

The first problem that arises in TTFA is the determination of the number of resolvable sources. This problem is not unique to TTFA but occurs in any application of an eigenvector analysis of data containing noise. In the absence of error, the eigenvalues beyond the true number of sources become zero except for calculational error. As will be explored later in this report, the choice becomes more difficult depending on the error in the data. Several approaches have been suggested in the literature (Duewar *et al.*, 1976; Hopke *et al.*, 1980). A large relative decrease in the magnitude of the eigenvalues is one indicator of the correct number of factors. In another test, the original data are reproduced using only the first factor and compared point-by-point with the original data. Several measures of the quality of fit are calculated including chi-squared

$$\chi^2 = \sum_{ij} \frac{(x_{ij} - \bar{x}_{ij})^2}{\sigma_{ij}^2} \tag{9}$$

where x_{ij} is the reconstructed data point using p factors and σ_{ij} is the measured uncertainty in the value of x_{ij}. The Exner function (Exner, 1966) is a similar measure and is calculated by

$$E^p = \left[\frac{\sum_{ij}(x_{ij} - \bar{x}_{ij})^2}{\sum_{ij}(x^o - \bar{x}_{ij})^2}\right]^{1/2} \tag{10}$$

where x^o is a grand ensemble average value. The empirical indicator function suggested by Malinowski (1977) is also calculated.

$$RSD = \left[\sum_{j=p+1}^{n} \frac{\lambda_j}{n(m-p)} \right]^{1/2} \quad \text{for } n > m \quad (11)$$

$$IND = \frac{RSD}{(n-p)^2} \quad (12)$$

$$RSD = \frac{\sum_{j=p+1}^{n} \lambda_j}{m(n-p)} \quad \text{for } n < m \quad (13)$$

$$IND = \frac{RSD}{(m-p)^2} \quad (14)$$

where λ_j are the eigenvalues from the diagonalization. This function has not proven as useful with particle composition data as Malinowski has found it to be with spectroscopy results (Malinowski, 1977). Finally the root-mean-square error and the arithmetic average of the absolute values of the point-by-point errors are also calculated. The data are next reproduced with both the first and second factors and again compared point-by-point with the original data. The procedure is repeated, each time with one additional factor, until the data are reproduced with the desired precision. If p is the minimum number of factors needed to adequately reproduce the data, then the remaining $n-p$ factors can be eliminated from the analysis. These tests do not provide unequivocal indicators of the number of factors that should be retained. Judgement becomes necessary in evaluating all of the test results and deciding upon a value of p. In this manner the dimension of the A and F matrices is reduced from n to p.

It is important to recognize that although the factor analysis allows a contraction of the dimensionality of the space to p axes, any set of eigenvectors that satisfy Equation (8) are mathematically acceptable. There are constraints that can be applied to limit the space to a smaller proportion of the full p dimensions. For example, if the columns of the A matrix are to represent physically real sources of airborne contaminants, then the elements of these vectors must be at least zero. There can be no negative values. If a sufficiently complete analysis has been made, the source profile column must then sum to 100%. However, this closure condition can cause other problems as described by Miesch (1976). For airborne particulate analysis, it is typical to have substantially less than 100% of the mass accounted for and this constraint is not applicable. Thus, there is a high degree of indeterminacy in the factor analysis solution because it has started with so little *a priori* information on the nature of the sources.

After the truncation of the A and F matrices to only p retained factors, it is necessary to interpret the nature of the factors for the system being studied. The A matrix produced by diagonalization cannot be directly associated with actual source profiles since it is one of an infinity of mathematically equivalent matrices that will diagonalize the correlation matrix. In order to relate this abstract matrix to one with physical significance, it is necessary to geometrically realign the factor axes of A with axes that represent real source emission concentration profiles. Such physically significant axes, called test vectors, are derived from existing knowledge of the relative elemental composition of actual source materials. The realignment procedure, called target transformation (Weiner *et al.*, 1970), involves finding a rotation vector, $\mathbf{r}$, which aligns a column of the A matrix with the input test vector, $\mathbf{b}$, by least-squares maximization of the overlap between a rotated axes of A and the test vector. Weighted least-squares has been found to greatly enhance the ability of the analysis to resolve sources with similar concentration profiles (Roscoe and Hopke, 1981a). The weighted target transformation rotation is given by

$$\mathbf{r} = (A^t W A)^{-1} A^t W \mathbf{b} \quad (15)$$

where W, the weight matrix, is a diagonal matrix that can have as its diagonal elements the inverse of the variance in the elemental concentrations or the inverse of the square of the average error of determination for the concentration values.

A suspected particle source can then be tested by determining if a factor axes can be rotated so as to overlap the suggested test vector. In addition to source profile test vectors, uniqueness test vectors that have all but one element set equal to zero and the remaining value set to unity are analyzed for all elements. An element is considered unique if its concentration in the data does not covary with the other measured species. The uniqueness test then determines if the concentration of an element is strongly related to any other element(s). The test vector predicted by the uniqueness test can then be used as a normalized source concentration profile for sources whose concentration profiles have not been measured.

It has been found the source profiles can be obtained by an iterative process from simple initial test vectors (Roscoe and Hopke, 1981a). The initial vector has a value of 1.0 for one element and zero for the others. The resulting rotation vector is used to obtain the new values for $\mathbf{b}'$.

$$\mathbf{b}' = A\mathbf{r}. \quad (16)$$

After setting any negative values to small positive ones, $\mathbf{b}'$ is used as the input to the next iteration. This process is repeated until the average per cent change in the values of $\mathbf{b}$ and $\mathbf{b}'$ are less than 10^{-4}. At this point these iterated profiles are available to be grouped p at a time to form a matrix B.

The quality of the least-squares fit of a factor axis to an input test vector is again dependent on the noise structure in the data. The least-squares method assumes that the error is symmetrically distributed and that may not be the case for source compositions. The

ability to screen known source profiles and to iteratively obtain source profiles will be tested as part of the results being reported here. It would be simple to test profiles if more information about samples or the sources were available. For example, Malinowski and McCue (1977) add artificial samples to their data describing pure components (100% contributed by a single known composition source) and can use these samples to guide the rotation. Similarly, if it were known that there was no contribution by a particular source to a particular sample, this information could be used to direct the analysis. However, the typical situation to which TTFA is applied is one in which this kind of knowledge of the system is unlikely to be available.

The sources that have been identified as contributing to the ambient concentrations are then used to calculate the corresponding mass contributions again using a least-squares method (Severin *et al.*, 1983).

$$\bar{\mathbf{f}} = (B^t W B)^{-1} B^t W \mathbf{x} \tag{17}$$

where $\bar{\mathbf{f}}$ is the mass contributions, B is the matrix of derived source profiles and $\mathbf{x}$ is a vector of the data matrix. The original data are then compared point-by-point with the reproduced concentrations to determine how well the chosen set of profiles represent the actual sources in the system. This comparison may suggest further refinements to the test vectors to improve the agreement. Again there are constraints on the solution in that the calculated mass contributions, f_{kj}, must be greater than or equal to zero. Calculations can be made with a number of possible combinations of vectors until the best p dimensional set is found. A computer code that performs all of these tasks has been described by Hopke *et al.* (1983), Hopke and Dharmavaram (1986) and Hopke (1988).

The absolute concentrations of the input source profile test vectors may not correspond exactly to those found in the system under study. Therefore, before the selected set of test vectors can be used to calculate the mass contribution of each source to each samples, they must be scaled to reflect the actual concentration of each element in the source material. The source vectors developed from the uniqueness test begin with one value set to a maximum concentration of unity and after iteration, therefore, require scaling to absolute concentration values. The scaling factors that can provide the correct calculated source contributions and the total measured sample weights are obtained by means of a multiple regression analysis (Hopke *et al.*, 1980) of the form

$$M_j = \sum_{k=1}^{p} s_k \bar{f}_{kj} \tag{18}$$

where M_j is the measured mass of sample j, $\bar{f}_{kj}$ is the mass matrix calculated in Equation 17 and s_k is the scaling factor to be calculated by the regression analysis. Comparison of the predicted and measured sample masses gives another indication of the quality of the fit with the selected set of test vectors. The coefficients should be positive and statistically significant. In addition when the source profiles are scaled, a_{ik}/s_k, the sum for a given source should be less than 1,000,000 ppm. Values > 1,000,000 have been seen and it indicates that too few sources were included in the analysis. These values can assist in determining if too many or too few sources have been included in the analysis.

Finally, the analysis is completed by estimating the uncertainties in the derived parameters. Since the error in the ambient concentration arises from error in both the source profiles and the mass contributions, it is necessary to estimate the errors in both A and F. However, there is no information available as to the fraction of error that should be apportioned to one matrix or the other. Unless there is some additional information available that describes the error structure of one or the other of these matrices, it is necessary to perform the analysis twice; once assuming A to be error-free and estimating the uncertainty in F and then the reverse case. Thus, an upper limit of error will be obtained for each of the variables. The error can be estimated in two ways that are described by Roscoe and Hopke (1981b). The more rigorous method involves a computationally intensive approach called jackknifing (Mosteller, 1971; Mosteller and Tukey, 1977). In this approach the parameters are calculated after eliminating one sample from the data set. When all possible subsets of data have been analyzed, the parameter values are averaged to yield a new set of estimates and the standard deviation of the distribution of parameter values becomes the uncertainty estimate.

It has been found that for most cases adequate estimates of these uncertainties can be obtained from a calculational approach described by Clifford (1973). It should be remembered that the errors obtained from this calculational method are going to be less precise estimators than those obtained with the jackknife method. However, this method yields estimates in a computationally easy manner and will suffice for most applications. It may be possible that with increasing computer capability that the more rigorous jackknife method will be more practical in the near future. However, it must be remembered that there is still the fundamental problem of partitioning the error between A and F that is irrelevant to the choice of estimation method.

Previous validation results

It is then necessary to test the TTFA system described above in terms of defining the true dimensionality of the system, developing accurate estimations of the actual source profiles for a system, and calculating accurate estimates of the mass contributions of the identified sources to the samples. A number of the previous reports on the analysis of actual ambient data have presented quite reasonable results. However, since the correct answers are not

known, it cannot be known if these results are accurate. Therefore, it becomes necessary to examine other data sets where comparisons can be made with the 'truth'. Hwang *et al.* (1984) have examined a simulated aerosol composition data set prepared as part of the Quail Roost II intercomparison study of receptor models (Currie *et al.*, 1984). Roscoe *et al.* (1984) have analyzed a set of elemental composition data for coal samples to determine the nature and amounts of the mineral phases present. They compare the TTFA results with independent estimates obtained using x-ray diffraction analysis of low temperature ashed samples. In both cases quite good agreement was obtained particularly considering no *a priori* information was used in completing the analysis. However, these data are not able to provide indications of the sensitivity of the analysis to variations in the error structure or the similarities in composition among the sources that contribute to the samples. Thus, to provide controlled sensitivity tests, several sets of synthetic data have been created.

Data set structure

In order to examine the effect of well defined error on the ability to determine the number of factors, to identify the proper sources and to accurately estimate the mass contributions of each source, several sets of data have been created. The first data sets are very simple two-component systems to examine the effect of error and orthogonality on the ability of TTFA to identify the dimensionality of the system and extract the source profiles. Ten variables were created with alternating values of 1 and 0 so the two vectors are orthogonal. Each variable was then multiplied by a uniformly distributed random number to examine the effects of column versus row scaling. The corresponding mass values for 40 'samples' were also obtained from a uniformly distributed random number generator. To test the system to determine the number of factors and iteratively develop the source profiles, two additional data sets were obtained by alternately adding and subtracting increments of 0.2 to the original 0,1 valued profiles. The data were then also perturbed by adding to each variable in the profile a normally distributed error before multiplying by the masses to yield the 'observed concentrations'. The final values were also perturbed by the addition of a normally distributed 'analytical' error. The data generation process can then be expressed as

$$a_{i1} = A_i(1-\Gamma) \qquad a_{i2} = A_i\Gamma$$
$$a_{i+1,1} = A_{i+1}\Gamma \qquad a_{i+1,2} = A_{i+1}(1-\Gamma) \quad (19)$$

where A_i are uniformly distributed random numbers between 0 and 1 and i has values 1, 3, 5, 7 and 9. These profiles are then rescaled such that

$$\sum_{i=1}^{10} a_{ik} = 1{,}000{,}000.$$

The data are then produced using the equation

$$x_{ij} = a_{i1} f_{ij} + a_{i2} f_{2j} + \varepsilon_{ij} \quad (20)$$

where the f values are also uniformly distributed random numbers for j equal to 1–40 periods, Γ has values of 0, 0.2 or 0.4, and ε_{ij} is a normally distributed error such that the average error in an x value is a given percentage. Thus, the data have a form similar to those described by Currie *et al.* (1984) or Watson *et al.* (1984). A second set of data can be generated by introducing an additional term to the equations for a_{i1} and a_{i2}. This term was normally distributed with a standard deviation equal to a given percentage of the value so that there is variation in the source composition profiles.

The second group of data sets were developed to look in more detail at just the effect of error on more realistic aerosol data sets, to test the ability of TTFA to identify the dimensionality of the problem, and to screen possible source profiles for inclusion in the source resolution. These sets were produced by recreating the first simulated data set originally developed by the National Bureau of Standards for the Quail Roost II Workshop on intercomparison of receptor methods (Currie *et al.*, 1984). Using the source profiles and mass values given in that report, an error-free data set is created to which random normally distributed error could be added with fixed standard deviations. These data sets were then subjected to TTFA analysis.

RESULTS AND DISCUSSION

The first task of the analysis is to identify the number of sources contributing to the particle mixture being sampled. This determination is affected by the analytical error, the collinearity of the sources, and the degree of variation in the source profiles. To illustrate these interactions, Table 1 shows the effect of collinearity and analytical error on the eigenvalues for the two component data sets described by Equation 19 with no variation in the source profile values. In the matrix diagonalization, the eigenvectors are always extracted in a way that maximizes the variance contained by the first vector. The second vector then contains the maximum amount of the remaining variance and so on for the additional vectors. As the source profiles move from being orthogonal, the projection of the second source vector on the first increases from zero as the cosine of the angle between them while the projection of the second vector on an orthogonal axis decreases as the sine of that angle. Thus, the total variance contained by the first eigenvector, its eigenvalue, increases. If the Γ value increases to 0.5, then there would be two identical sources and there would be only a single significant eigenvalue. Thus, the degree to which the sources are similar directly affects the values of the second eigenvalues relative to the first.

The second effect included in Table 1 is the analytical error. It should be noted that as the analytical

Table 1. Effect of collinearity (Γ) and analytical error on Q-mode eigenvalues

Factor	0%			1%		
	0.0	0.2	0.4	0.0	0.2	0.4
1	3.34E+01	3.70E+01	3.96E+01	3.34E+01	3.70E+01	3.96E+01
2	6.59E+00	2.96E+00	3.73E−01	6.58E+00	2.95E+00	3.68E−01
3	1.04E−11	1.15E−11	1.48E−11	9.28E−04	9.02E−04	8.71E−04
4	6.74E−12	8.73E−12	9.37E−12	4.90E−04	4.64E−04	4.54E−04
5	6.14E−12	6.51E−12	8.04E−12	3.75E−04	3.73E−04	3.72E−04

Factor	3%			5%		
	0.0	0.2	0.4	0.0	0.2	0.4
1	3.34E+01	3.70E+01	3.73E+01	3.37E+01	3.70E+01	3.96E+01
2	6.56E+00	2.93E+00	8.37E−01	6.24E+00	2.91E+00	3.60E−01
3	8.44E−03	8.19E−03	7.29E−01	1.90E−02	2.29E−02	2.19E−02
4	4.37E−03	4.14E−03	3.55E−01	1.00E−02	1.14E−03	1.13E−02
5	3.39E−03	3.37E−03	3.38E−01	6.41E−03	9.41E−03	9.28E−03

uncertainty increases, the higher order eigenvalues increase and the difference between the second and third eigenvalues decreases. It would be possible to increase the uncertainty to the point where the values of the true variance due to variation in the source strength is about the same as the contribution from the noise. Thus, analytical precision is directly related to the ability of a factor analysis to properly identify the number of independent sources.

A number of other indicators have been suggested, particularly those related to data reproduction. In the case of only collinearity and analytical error, these parameters, RMS error, Chi-Squared, the Exner functions and the average absolute point-by-point error, become small at two factors and remain small for increasing number of factors. They are affected by the addition of error and the collinearity. Some examples are shown in Table 2. Although the values for factors 3 and higher do increase with increasing analytical error, there is still sufficient discrimination in these cases to identify 2 factors. It is clear that the collinearity has a very significant effect on the relative values of these parameters.

To examine the effect of variation in the source profiles, several sets of data with Γ of 0.2 were created in which a normally distributed 1% variation was added to the a values before calculating the concentration values. Normally distributed analytical error was then added to the total x_{ij} value. In the case of zero analytical errors, the values are very similar to those given in Table 2. For 1, 3 and 5% analytical error, the dimensionality tests are given in Table 3. It can be seen that the variation in the profiles causes similar effects to those of the collinearity and an analytical error.

In the analysis of the recreated NBS data sets, analytical errors of 1, 3, 5 and 10% were added. The resulting Q-mode eigenvalues are intercompared in Fig. 1. With error-free data, it is clear there are nine sources. Because of the collinearity that exists in the data, only five of them account for most of the sample variance. Henry (1983) has suggested that there are seven resolvable sources although several of these are linear combinations of the original sources. In addition of analytical error causes the curve to rapidly flatten and it becomes difficult to distinguish weak factors from the noise. The other dimensionality measures act similarly and the choice of factor number becomes equivocal but in the range of 5 to 7.

The next test is to examine the ability to derive the correct profiles. The results for the iterative derivation of one of the two possible vectors of the simple two vector data set of Equation 19 are shown in Table 4. It can be seen that when the original vectors are fully orthogonal, the iterative procedure correctly finds the two vectors. However, as the vectors become more collinear, the iterative procedure still derives a pair of orthogonal axes. However, in Table 5, the results of the iterative procedure are given for the data set with $\Gamma = 0.2$, analytical error of 1% and the addition of normally distributed variation in the profiles. In this set, the iterative procedure is able to move from the orthogonal axes and obtain profiles that are quite like the original values. The reduction of the profile variation to 0.1% and the retention of 1% analytical error causes the process to once again derive the orthogonal vectors. At 0.1% profile variation and no error, it again properly finds the oblique vectors. However, if the profile variation is reduced to 0.01% with no error, the orthogonal solution is again obtained. Thus, it is the variation in source profile that permits the iterative TTFA approach to obtain the correct oblique solution.

To further study the effect of collinearity, Γ is increased to 0.4. A solution (not shown) that is close to correct is only obtained with 1% profile variation and less than 1% analytical error. With 3% analytical error added, the solution becomes very similar to that of the $\Gamma = 0.2$ data. Thus, the iterative procedure can

Table 2. Effects of collinearity and analytical error on dimensionality tests

Factor	0% Error			1% Error			3% Error			5% Error		
	0.0	0.2	0.4	0.0	0.2	0.4	0.0	0.2	0.4	0.0	0.2	0.4
						RMS error						
1	5.94E−02	3.58E−02	1.19E−02	5.92E−02	3.56E−02	1.19E−02	5.91E−02	3.57E−02	1.23E−02	5.95E−02	3.59E−02	1.34E−02
2	1.30E−07	1.30E−07	1.34E−07	1.36E−03	1.30E−03	1.26E−03	4.07E−03	3.89E−03	3.55E−03	6.41E−03	6.50E−03	6.29E−03
3	1.17E−07	1.18E−07	1.21E−07	1.07E−03	1.03E−03	1.01E−03	3.21E−03	3.10E−03	2.75E−03	4.74E−03	5.17E−03	5.06E−03
4	1.05E−07	1.03E−07	1.06E−07	8.29E−04	8.08E−04	8.06E−04	2.49E−03	2.43E−03	2.25E−03	3.87E−03	4.05E−03	4.06E−03
5	9.46E−08	8.91E−08	9.29E−08	6.26E−04	6.17E−04	6.24E−04	1.88E−03	1.85E−03	1.62E−03	3.08E−03	3.08E−03	3.09E−03
						Exner function						
1	0.6644	0.4980	0.1941	0.6641	0.4974	0.1934	0.6642	0.4974	0.1974	0.6670	0.4988	0.2050
2	0.0000	0.0000	0.0000	0.0127	0.0141	0.0153	0.0382	0.0423	0.0441	0.0596	0.0705	0.0764
3	0.0000	0.0000	0.0000	0.0099	0.0110	0.0121	0.0298	0.0332	0.0345	0.0450	0.0553	0.0602
4	0.0000	0.0000	0.0000	0.0081	0.0091	0.0100	0.0244	0.0274	0.0279	0.0369	0.0458	0.0499
5	0.0000	0.0000	0.0000	0.0064	0.0072	0.0079	0.0192	0.0217	0.0217	0.0307	0.0362	0.0349
						Average error						
1	101.48	28.55	8.14	101.07	28.50	8.16	100.46	28.53	8.64	101.88	28.65	8.95
2	0.00	0.00	0.00	0.81	0.73	0.72	2.41	2.19	2.23	4.05	3.66	3.63
3	0.00	0.00	0.00	0.71	0.64	0.63	2.11	1.91	1.99	3.94	3.19	3.16
4	0.00	0.00	0.00	0.70	0.55	0.54	2.11	1.66	1.72	5.20	2.77	2.73
5	0.00	0.00	0.00	0.72	0.49	0.46	2.15	1.48	1.44	4.76	2.48	2.32

Table 3. Effects of variation in profiles on dimensionality tests

(Q-Mode, Gamma = 0.2, Error = 1%)

Factor	Without error in profiles Eigenvalue	RMS	EXNER	Avg. error	With error in profiles Eigenvalue	RMS	EXNER	Avg. error
1	3.70E+01	3.57E−02	0.4974	28.50	3.70E+01	3.57E−02	0.4980	31.31
2	2.95E+00	1.30E−03	0.0141	0.73	2.95E+00	1.28E−03	0.0140	2.85
3	9.02E−04	1.03E−03	0.0110	0.64	9.60E−04	9.62E−04	0.0107	2.64
4	4.64E−04	8.08E−04	0.0091	0.55	5.02E−04	7.76E−04	0.0085	2.56
5	3.73E−04	6.17E−04	0.0072	0.49	2.99E−04	5.60E−04	0.0069	2.40

(Q-Mode, Gamma = 0.2, Error = 3%)

Factor	Without error in profiles Eigenvalue	RMS	EXNER	Avg. error	With error in profiles Eigenvalue	RMS	EXNER	Avg. error
1	3.70E+01	3.57E−02	0.4974	28.53	3.70E+01	3.57E−02	0.4983	29.03
2	2.93E+00	3.90E−03	0.0423	2.19	2.95E+00	3.66E−03	0.0400	2.29
3	8.19E−03	3.10E−03	0.0332	1.91	7.90E−03	2.76E−03	0.0306	2.05
4	4.14E−03	2.43E−03	0.0274	1.66	3.83E−03	2.25E−03	0.0248	1.90
5	3.37E−03	1.85E−03	0.0217	1.48	2.48E−03	1.66E−03	0.0202	1.68

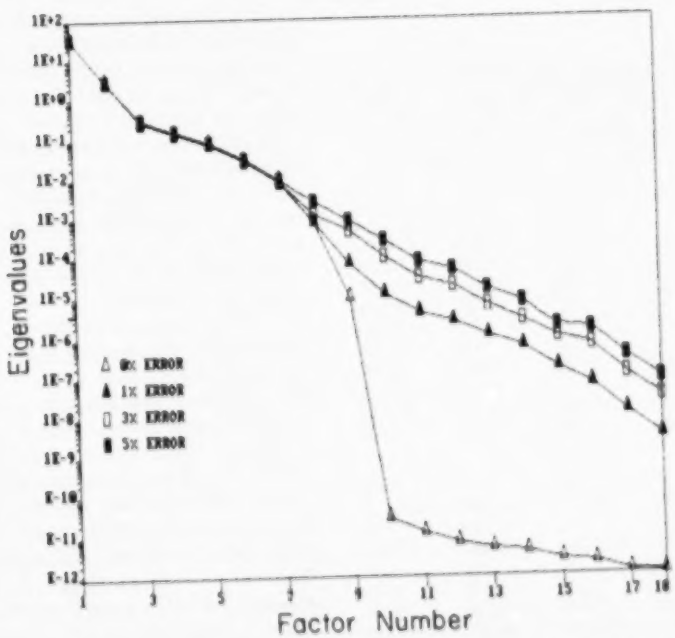

Fig. 1. Eigenvalues for the recreated NBSI data set with various levels of analytical error.

track nonorthogonal profiles if there is variability of the profiles. The addition of error reduces the ability of the procedure to find the proper axes. As collinearity increases, there is a point dependent on both source variability and measurement error beyond which the correct source profiles cannot be found. This complex interaction of the collinearity, the source profile variability and the analytical precision reduces the sensitivity of the iterative rotation for the determination of nonorthogonal axes and will take additional extensive efforts to define fully.

To examine this question further, the first of the original N.B.S. data sets described by Currie *et al.* (1984) has been reexamined. In this data set, there are substantial collinearities between the sources. However, for this data set, the nine source profiles are fixed

Table 4. Effects of analytical error on the iterative development of source profiles

Truth	0%	1%	3%	5%
Gamma = 0.0				
201,075	201,075	208,015	207,389	203,643
0	0	934	2,729	14,469
272,582	272,582	282,021	281,318	275,410
0	0	180	754	3,224
157,261	157,261	162,807	162,579	159,388
0	0	0	36	71
95,563	95,563	98,998	98,955	94,885
0	0	0	0	4,142
273,519	273,519	246,759	245,341	238,783
0	0	287	898	5,986
Gamma = 0.2				
181,073	208,278	208,087	207,715	207,209
74,169	38	1,146	3,173	4,911
245,450	282,340	281,767	280,815	279,750
2,321	0	497	1,071	2,445
141,628	162,915	162,800	162,601	162,213
507	0	0	38	53
86,069	99,018	98,941	98,907	98,851
23,839	0	0	0	4
215,095	247,410	246,303	244,227	242,056
29,847	0	459	1,453	2,508

Table 5. Effects of variations in the profile on the iterative development of source profiles (Gamma = 0.2, Error = 1%)

	Without variation in profiles		With variation in profiles	
Truth	*Q*	*R*	*Q*	*R*
Profile 1				
177,367	208,079	207,536	156,253	171,391
72,654	1,142	1,460	122,618	86,930
240,429	281,768	281,989	211,777	232,364
22,749	490	473	38,280	27,045
138,717	162,791	162,991	122,255	134,343
479	10	15	753	501
84,305	98,938	99,022	74,015	81,350
23,363	0	1	39,451	27,840
210,689	246,312	245,991	185,134	203,205
29,249	470	523	49,462	35,031
Profile 2				
54,957	21	342	8,541	5,400
360,184	487,337	487,676	469,409	475,960
74,496	1,014	320	11,568	7,739
112,779	152,647	152,684	147,407	149,414
42,981	311	8	6,169	4,006
2,373	3,203	3,201	3,828	3,785
26,122	323	283	4,086	2,514
115,823	157,222	157,215	151,067	152,825
65,281	1,996	2,780	9,532	6,633
145,005	195,926	195,491	188,393	191,724

and there are analytical errors of the order of 10% imposed on the resulting concentrations. In the previous two component cases, there was little difference observed between the *Q*- and *R*-mode derived profiles.

In the NBS data set, there are some differences. For example, Table 6 shows the *Q*- and *R*-mode profiles iterated from the unique vector for carbon. These profiles are compared to the Wood Smoke profile. It can be seen that the *R*-mode profile is very much closer to the true profile and the *Q*-mode profile contains very much larger silicon and lower potassium values.

Alternatively, Table 6 presents results for the *Q*- and *R*-mode vectors iterated from the unique vectors that best correspond with the Soil source. Here the *Q*-mode vector may be preferable because it does not have the large erroneous carbon value shown by the *R*-mode vector. Similarly, the *Q*-mode vector also shown in Table 6 is a better representation of Sandblast than is the *R*-mode vector. Also given in Table 6 are intercomparisons on the Road profile with the best iterated profiles. Again the *Q*-mode profile is in better agreement with the actual profile.

There are similar results for other profiles with most being better represented by the *Q*-mode vectors and only a few by the *R*-mode. It appears that for profiles that are dominated by a single element (C in Wood Smoke, for example) are better represented by the *R*-mode vectors.

Those profiles that have the same major elements in common appear to be better developed from the *Q*-mode vectors. In some cases a reasonable profile could be derived only from one mode vectors. For example, a reasonable Basalt profile could be obtained from the *Q*-mode vectors but not from the *R*. For Oil-B, the *R*-mode was much better but still substantially underestimated the carbon concentration. This result presumably is due to the uniqueness of the Wood Smoke profile for C making it difficult to find C in the other vectors. It appears the best approach to solving a particular airshed problem is to derive both sets of profiles before trying to find the best combination of vectors to fit the data.

A question arises as to why the iterative procedure works for this data set since the profiles are fixed and the simple two vector results discussed previously seemed to show that it was the profile variation that drives the iterative development. It is suggested that the collinear sources that exist in this data set such as Soil and Basalt have different proportions in different samples. These two sources are unresolved and thus represent an average profile that has a varying composition. Those profiles that are essentially orthogonal to the others, i.e. Wood Smoke, can be found in the absence of profile variation because of their orthogonality as seen with the two component test data. A high value for a single (or very few) element(s) may cause the vector to be sufficiently orthogonal that it is found without great difficulty.

Although the iterative procedure is unable to develop the profiles in the case of strong collinearity and/or high noise, the target transformation rotation can also be used to screen source profiles and to identify sources that may be useful to include in other analyses such as a CMB approach. By screening it is meant that the source profile is used as the input test vector in Equation 15 and the reproduced vector calculated using Equation 16. The ability to identify

Table 6. Comparison of *R*- and *Q*-mode iterated unique vectors with original N.B.S. profiles

Element	Wood: Rescaled original profile	Wood: Iterated *R*-mode vector	Wood: Iterated *Q*-mode vector	Element	Sandblast: Rescaled original profile	Sandblast: Iterated *R*-mode vector	Sandblast: Iterated *Q*-mode vector
C	924,741	905,250	740,140	C	8,812	1,072	3,995
Na	4,624	10,415	3,417	Na	14,100	0	24,399
Al	10,789	172	0	Al	13,219	162,420	0
Si	13,717	63,538	237,640	Si	704,995	571,990	770,020
S	1,541	4,803	0	S	15,510	34,646	12,066
Cl	15,412	2	7,147	Cl	49,350	1,874	43,033
K	12,329	683	5	K	12,337	8,386	2,511
Ca	13,871	3,221	0	Ca	14,452	46,873	12,297
Ti	31	0	310	Ti	7,050	17,437	7,721
V	8,665	20	135	V	53	7	1
Cr	2	468	253	Cr	10,222	914	7,373
Mn	31	894	643	Mn	2,467	2,184	1,748
Fe	1,541	7,342	8,889	Fe	111,037	139,600	89,162
Ni	15	2	23	Ni	5,640	482	4,320
Cu	53,502	372	0	Cu	12,161	345	8,586
Zn	1,079	836	745	Zn	8,284	0	6,047
Br	123	1,899	653	Br	88	434	51
Pb	154	88	5	Pb	10,222	11,336	6,666

Element	Road: Rescaled original profile	Road: Iterated *R*-mode vector	Road: Iterated *Q*-mode vector	Element	Soil: Rescaled original profile	Soil: Iterated *R*-mode vector	Soil: Iterated *Q*-mode vector
C	486,466	75,260	354,290	C	67,644	189,510	35,734
Na	59,533	134,720	85,829	Na	4,075	59,363	24,970
Al	70,305	309,870	86,646	Al	226,159	223,780	256,470
Si	166,691	249,420	0	Si	553,173	438,110	540,560
S	6,917	4,806	55,644	S	306	9	12,594
Cl	9,185	125	18,964	Cl	102	2,270	4,337
K	5,670	7,962	1,459	K	12,225	14,923	10,057
Ca	15,762	21,655	43,020	Ca	20,884	11,584	7,604
Ti	3,629	1	7,112	Ti	12,938	5,765	11,755
V	113	1,411	6	V	306	833	5
Cr	113	25	433	Cr	306	152	818
Mn	1,021	1,518	968	Mn	2,037	1,916	2,292
Fe	33,452	5,548	56,165	Fe	98,817	46,660	88,584
Ni	79	573	453	Ni	102	216	301
Cu	567	403	506	Cu	204	19	190
Zn	2,835	3,158	5,350	Zn	407	560	1,209
Br	35,493	52,889	70,272	Br	10	4,333	934
Pb	102,169	130,660	212,890	Pb	306	1	1,584

the presence of specific vectors in a data set was, therefore, tested using the NBS data set. The reproduced values for two selected profiles are given in Table 7. It can be seen that the profiles are easily distinguished as belonging or not belonging to the source matrix when there is no error in the data. However, the screening process becomes more difficult as the level of error increases. There are very significant decreases in the quality of the vector reproduction as the error in the data increases. However, for elements for which the particular profile is the dominant source, such as Zn from Incinerator or C from Wood Smoke, the values stay quite constant.

To show results for the entire profile set, Table 8 presents the average per cent differences between the predicted vector produced by TTFA and the original profile for each of the 14 profiles either provided to the Quail Roost II participants as possible sources as well as the ones actually used. Sandblast had not been provided and Road was a composite of 75% motor vehicles and 25% soil.

From the size of the error values in Table 8, it can be clearly seen which sources should be used when there is no error. However, the addition of uncertainty rapidly raises the error in the source profile fits. In addition to the created NBS data with fixed error

Table 7. Effects of analytical error on the screening of source profiles by TTFA

Incinerator

Element	Original profile	0%	1% Error		3% Error		5% Error		Original data set	
			Q	*R*	*Q*	*R*	*Q*	*R*	*Q*	*R*
C	40.0	40.0	35.0	43.0	30.0	50.0	−18.3	54.1	60.9	45.0
Na	120	120	111	117	109	115	109	111	184	141
Al	4.2	4.2	0.0	0.0	0.0	0.0	−58.9	−12.2	65.5	8.7
Si	7.0	7.0	0.0	0.0	0.0	0.0	−10.3	−185	−250	−67.6
S	130	130	132	126	130	124	117	125	189	173
Cl	270	270	277	273	282	275	291	276	379	361
K	200	200	202	202	203	202	207	202	293	255
Ca	11.0	11.0	9.6	9.4	10.0	8.9	13.5	7.5	2.8	10.6
Ti	0.90	0.90	0.66	0.88	0.79	1.0	1.1	0.76	−2.8	−0.11
V	0.03	0.03	0.06	0.11	0.067	0.14	0.10	0.21	0.63	0.052
Cr	0.9	0.9	0.84	0.98	0.81	1.1	0.72	1.0	0.24	0.43
Mn	0.30	0.30	0.63	0.31	0.66	0.25	0.65	0.25	1.4	0.92
Fe	6.1	6.1	6.3	16.0	8.0	2.2	17.8	45.0	7.1	9.9
Ni	0.1	0.1	0.066	0.12	0.0	0.1	−0.25	0.087	0.4	0.29
Cu	3.6	3.6	3.7	3.7	4.0	3.9	4.4	4.1	6.6	4.3
Zn	26.0	26.0	25	26.0	24.0	25.0	22.1	24.7	39.2	35.1
As	0.20	0.20	0.20	0.19	0.20	0.18	0.22	0.17	—	—
Br	0.80	0.80	3.2	4.5	5.0	6.8	9.94	12.2	2.7	1.1
Pb	17.0	17.0	23.6	12.4	29.0	10.0	37.5	7.2	19.7	21.6
Average % deviation:		0.0812	70	104	103	165	221	317	331	84

Wood Smoke

Element	Original profile	0%	1% Error		3% Error		5% Error		Original data set	
			Q	*R*	*Q*	*R*	*Q*	*R*	*Q*	*R*
C	600	600	594	589	588	578	559	549	589	583
Na	3	3.0	0.85	1.27	−0.56	1.3	−4.4	0.091	7.9	3.9
Al	7.0	7.0	5.3	2.93	3.4	0.76	−6.27	−9.43	−40.1	−12.7
Si	8.9	9.0	−14.8	−16.4	−19.9	−19.8	−26.3	−9.44	30.2	−34.6
S	1.0	0.98	3.3	4.32	2.9	4.6	−0.33	3.6	0.35	3.9
Cl	10.0	10.0	9.6	9.6	10.3	7.9	11.4	5.2	9.7	7.8
K	8.0	8.0	6.8	6.6	6.4	6.1	6.2	5.5	6.8	6.5
Ca	9.0	9.0	8.0	8.0	7.8	7.7	7.9	8.3	6.1	4.6
Ti	0.02	0.02	0.085	−0.33	0.26	−0.44	0.6	−0.59	−0.59	−0.37
V	0.0	0.00036	−0.0077	−0.014	−0.015	−0.025	−0.014	−0.041	0.11	−0.075
Cr	0.0	0.00049	−0.04	−0.086	−0.087	−0.14	−0.25	−0.27	−0.032	0.098
Mn	0.02	0.019	0.21	0.26	0.27	0.31	0.37	0.34	0.23	0.47
Fe	1.0	1.0	0.67	4.8	0.29	7.8	0.65	13	3.2	0.66
Ni	0.01	0.01	0.052	0.0037	0.06	−0.022	0.058	0.091	−0.023	−0.12
Cu	0.0	0.00051	−0.13	−0.053	−0.15	−0.05	−0.083	0.012	−0.15	0.17
Zn	0.70	0.70	0.94	0.91	0.88	1.1	0.69	1.13	1.0	1.3
As	0.0	−0.00006	0.0011	0.0012	0.0042	0.0043	0.012	0.011	—	—
Br	0.08	0.076	0.20	−0.55	0.60	−0.59	2.19	0.35	0.23	0.43
Pb	0.1	0.091	0.56	1.7	2.2	2.2	7.8	5.4	−0.35	−1.1
Average % deviation:		1.29	158	347	342	457	870	707	363	473

Table 8. Effect of error on the screening of source profiles

Source	0% Error	1% Error		Original data	
		Q	*R*	*Q*	*R*
Steel-A*	21.2	3941	1566	1599	1264
Steel-B	641	1888	2336	1108	1509
Oil-A	3857	2462	4941	1611	2852
Incinerator*	0.081	70	104	331	84
Oil-B*	0.081	20.4	25	50	80
Coal-A	215	222	174	36	35
Aggregate	1217	1008	984	159	233
Sandblast*	0.414	20.3	53	71	36
Glass	3579	3773	4425	1074	1176
Basalt*	4.63	1023	1339	1389	98
Soil*	1.05	1143	201	68	227
Wood*	1.29	158	347	363	473
Coal-B*	1.37	5.18	700	220	110
Road*	0.0088	3.1	280	30	23

* Sources actually used to create the data set.

levels, the same procedure was applied to the original data set. The original data set had varying levels of error instead of a fixed level for all variables and thus they behave somewhat differently. There are also differences between the *R*- and *Q*-mode analyses. The process could be used to eliminate some obvious cases such as Glass, Steel-B and Oil-A. However, it cannot properly distinguish between sources like Coal-A and Coal-B in the actual data set. It would clearly suggest that Steel-A does not belong even when it was used to create the data set. However, this source contributes only insignificantly to the average mass concentration because it was created with a very low emission rate. Thus, this screening process may help to decide which sources contribute to the samples, but at typical levels of analytical precision, difficulties remain in making accurate assignments.

An alternative approach to screening the profiles for possible use would be to use them as the input vector to the iterative process. The process can then determine if the resulting profile is sufficiently unaltered so as to suggest that it does, in fact, belong to the set of contributing sources. Table 9 presents the results of iterating on several of the starting vectors. Again some results are quite clear. The Road source gives back a profile quite close to the original one. Several of the other profiles actually used behave similarly including Sandblast, Oil-B and Incinerator. The average per cent errors for these vectors are below 50% and they are quite clearly identified as belonging to the set. Coal-B shows values of just below 100% and thus still appears to belong. The *R*-mode profile for Wood Smoke looks fairly close. However, the small values for most of the elements cause small absolute deviations to give large percentage differences. Soil and Steel-A yield values of a few hundred per cent average deviation. In the case of Soil, it appears that a linear combination of Soil and Basalt is being obtained. Because the mass contributions of these area sources are correlated by the way the data set was created, this result is to be expected. Hwang *et al.* (1984) have shown that the derived Soil/Basalt source accounted very well for the sum of the two actual sources.

It can be seen from Table 9 that neither Basalt nor Glass are well reproduced. Similar results are obtained for Oil-A and Steel-B. Somewhat equivocal results are obtained for Aggregate and Coal-A where the iterated profiles strongly resemble those derived from Sandblast and Coal-B, respectively. However, it appears that it is feasible to use the iterative process to test suspected source profiles for use in the source apportionment.

CONCLUSIONS

It is clear from these results that TTFA will not provide unique solutions to source apportionment problems. Although previous studies have shown that quite reasonable agreement can be obtained with known results, there are limitations to the ability to identify the number of sources and to derive their profiles. There are complex interactions among the collinearity of the source profiles, the variation in those profiles and the analytical precision in the data that cause a blurring of the distinction between weak but truly contributing sources and factors that arise from the noise. It is likely that this analysis will typically underestimate the true number of sources and will tend to identify linear combinations of similar profiles. It can find profiles that are naturally orthogonal to the other profiles. The iterative development of source profiles appears to depend on variation in those profiles to find them but the resolution is degraded as the analytical precision becomes poorer. Finally it seems that there is no way to remove the need for operator judgement in choosing the number of factors and the selection of profiles to include in fitting the data. The technique can help to identify unusual sources that may not have been previously considered and to find the major contributing source classes. It can be applied without direct knowledge of the specific compositions of sources in an airshed. It can thus be used to analyze existing data to assist in planning further, more detailed studies so that definitive management strategies can be developed.

Acknowledgements—This work was supported in part by the U.S. Environmental Protection Agency under contract 68-02-3449 and the Nationaal Fonds voor Wetenschappelijk Onderzoek (Belgium). Although the work has been supported by the Environmental Protection Agency, it has not been reviewed by the Agency and no official endorsement of this report should be inferred. The author would like to express his gratitude to Professor D. L. Massart and Dr L. Kaufman of the Vrije Universiteit Brussel for their hospitality and the facilities to complete the studies presented in this report.

Table 9. Results of iterations starting from the original profiles

Element	Steel-A Original vector	Steel-A Iterated *R*-mode vector	Steel-A Iterated *Q*-mode vector	Element	Road Original vector	Road Iterated *R*-mode vector	Road Iterated *Q*-mode vector
C	300,842	201,590	75,120	C	486,466	480,460	525,130
Na	22,062	42,518	21,174	Na	59,533	47,035	49,560
Al	26,073	225,140	18	Al	70,305	88,086	62,500
Si	16,045	337,460	677,190	Si	166,691	169,660	116,430
S	54,152	382	24,039	S	6,917	7,575	11,457
Cl	60,168	25,943	34,094	Cl	9,185	10,395	12,698
K	18,251	28,986	520	K	5,670	4,040	4,513
Ca	140,393	48,257	25,814	Ca	15,762	17,417	24,780
Ti	2,607	7,666	8,410	Ti	3,629	3,497	4,417
V	3,610	831	161	V	113	93	2
Cr	6,619	1,889	6,466	Cr	113	232	221
Mn	32,090	2,220	1,640	Mn	1,021	1,073	980
Fe	240,674	65,040	93,503	Fe	33,452	31,306	37,613
Ni	5,616	1,164	3,831	Ni	79	139	185
Cu	5,215	3,005	7,590	Cu	567	157	42
Zn	50,140	4,541	4,995	Zn	2,835	2,942	3,146
Br	201	3,305	2,286	Br	35,493	35,125	36,665
Pb	15,243	68	13,150	Pb	102,169	100,770	109,670
Average % deviation:		307.7	348.5	Average % deviation:		22.4	40.5

Element	Basalt Original vector	Basalt Iterated *R*-mode vector	Basalt Iterated *Q*-mode vector	Element	Glass Original vector	Glass Iterated *R*-mode vector	Glass Iterated *Q*-mode vector
C	192	3,983	35,371	C	5,250	18,521	25,526
Na	36,539	2	1	Na	393,783	160,901	211,283
Al	169,234	311,660	247,660	Al	9,451	157,511	48,468
Si	461,574	345,870	392,310	Si	6,563	278,571	416,377
S	677	10,613	4,623	S	446,288	252,517	199,613
Cl	385	24	1,824	Cl	7,088	20,065	18,979
K	15,962	24,269	28,984	K	70,881	20,718	20,565
Ca	128,849	107,300	99,033	Ca	12,076	12,964	8,123
Ti	17,308	21,897	18,658	Ti	1,050	17	7,085
V	385	329	667	V	3,938	1,795	1,023
Cr	192	271	243	Cr	8,926	2,332	2,604
Mn	2,885	3,196	2,797	Mn	788	311	56
Fe	165,388	164,610	160,120	Fe	26,252	60,779	32,333
Ni	192	195	299	Ni	3,413	2,307	1,606
Cu	192	649	731	Cu	1,050	1,778	1,198
Zn	154	1,246	855	Zn	1,838	2,064	998
Br	55	175	528	Br	53	6,084	4,149
Pb	19	3,720	5,297	Pb	1,313	765	10
Average % deviation:		1364.2	2727.3	Average % deviation:		1021.7	903.7

REFERENCES

Alpert D. J. and Hopke P. K. (1980) A quantitative determination of sources in the Boston urban aerosol. *Atmospheric Environment* **14,** 1137–1146.

Alpert D. J. and Hopke P. K. (1981) A determination of the sources of airborne particles collected during the regional air pollution study. *Atmospheric Environment* **15,** 675–687.

Blifford I. H. and Meeker G. O. (1967) A factor analysis model of large scale pollution. *Atmospheric Environment* **1,** 147–157.

Bogen J. (1973) Trace elements in the atmospheric aerosol in the Heidelberg area measured by instrumental neutron activation analysis. *Atmospheric Environment* **7,** 1117–1125.

Clifford A. A. (1973) *Multivariate Error Analysis.* Applied Science Publishers, London.

Curran T. C. and Suggs J. C. (1986) Effects of measurement uncertainty on air quality summary statistics. *Atmospheric Environment* **20,** 571–576.

Currie L. A., Gerlach R. W., Lewis C. W., Balfour W. D., Cooper J. A., Dattner S. L., DeCesar R. T., Gordon G. E., Heisler S. L., Hopke P. K., Shah J. J., Thurston G. D. and Williamson H. J. (1984) Interlaboratory comparison of source apportionment procedures: results for simulated data sets. *Atmospheric Environment* **18,** 1517–1537.

Duewer D. L., Kowalski B. R. and Fasching J. L. (1976) Improving the reliability of factor analysis of chemical data by utilizing the measured analytical uncertainty. *Anal. Chem.* **48,** 2002–2010.

Dzubay T. G. (1980) Chemical element balance method applied to dichotomous sampler data. *Annals N. Y. Acad. Sci.* **338,** 126–144.

Exner O. (1966) Additive physical properties. *Collection of Czech Chem. Commun.* **31,** 3222–3253.

Friedlander S. K. (1973) Chemical element balances and identification of air pollution sources. *Envir. Sci. Technol.* **7,** 235–240.

Gaarenstroom P. D., Perone S. P. and Moyers J. L. (1977) Application of pattern recognition and factor analysis for characterization of atmospheric particulate composition in southwest desert atmosphere. *Envir. Sci. Technol.* **11,** 795–800.

Gatz D. F. (1975) Relative contributions of different sources of urban aerosols: application of a new estimation method to multiple sites in Chicago. *Atmospheric Environment* **9,** 1–18.

Gatz D. F. (1978) Identification of aerosol sources in the St. Louis area using factor analysis. *J. appl. Met.* **17,** 600–608.

Gladney E. S. (1974) Trace Elemental Emissions from Coal-Fired Power Plants: A Study of the Chalk Point Electric Generating Station. Ph.D. Thesis, University of Maryland, College Park, MD.

Greenberg R. R. (1976) A Study of Trace Elements Emitted on particles from Municipal Incinerators. Ph.D. Thesis, University of Maryland, College Park, MD.

Heidam N. Z. and Kronberg D. (1985) Comments on "A Comparison of *R*- and *Q*- Modes in Target Transformation Factor Analysis for Resolving Environmental Data". *Atmospheric Environment* **19,** 1549–1551.

Heindryckx R. and Dams R. (1974) Continental, marine, and anthropogenic contributions of the aerosol of an individual zone. *J. Radioanal. Chem.* **19,** 339–349.

Henry R. C. (1983) Stability analysis of receptor models that use least-squares fitting. In *Receptor Models Applied to Contemporary Pollution Problems* (edited by Dattner S. L. and Hopke P. K.) pp. 141–157. Air Pollution Control Association, Pittsburgh, PA.

Hopke P. K. (1980) Source identification and resolution through application of factor and cluster analysis. *Annals N. Y. Acad. Sci.* **338,** 103–115.

Hopke P. K. (1985) *Receptor Modeling in Environmental Chemistry*. J. Wiley & Sons, New York.

Hopke P. K. (1988) *FANTASIA: Factor Analysis to Apportion Sources in Analyses, User's Manual*. Elsevier Science Publishers, Amsterdam.

Hopke P. K. and Dharmavaram S. (1986) Recent improvements to FANTASIA, a target transformation factor analysis program. *Comput. Chem.* **10,** 163–164.

Hopke P. K., Alpert D. J. and Roscoe B. A. (1983) FANTASIA—a program for target transformation factor analysis to apportion sources in environmental samples. *Comput. Chem.* **7,** 149–155.

Hopke P. K., Gladney E. S., Gordon G. E., Zoller W. H. and Jones A. G. (1976) The use of multivariate analysis to identify sources of selected elements in the Boston urban aerosol. *Atmospheric Environment* **10,** 1015–1025.

Hopke P. K., Lamb R. E. and Natusch D. F. (1980) Multielemental characterization of urban roadway dust. *Envir. Sci. Technol.* **14,** 164–172.

Hwang C. S., Severin K. G. and Hopke P. K. (1984) Comparison of *R*- and *Q*-mode factor analysis for aerosol mass apportionment. *Atmospheric Environment* **18,** 345–352.

Kowalczyk J. F., Core J. E. and Cooper J. A. (1979) The true sources of Portland's suspended particulate. Paper No. 79-10.5, Air Pollution Control Association, Pittsburgh, PA.

Kowalczyk G. S., Gordon G. E. and Rheingrover S. W. (1982) Identification of atmospheric particulate sources in Washington, D.C. using chemical element balances. *Envir. Sci. Technol.* **16,** 79–90.

Liu C. K., Roscoe B. A., Severin K. G. and Hopke P. K. (1982) The application of factor analysis to source apportionment of aerosol mass. *Am. Ind. Hyg. Assoc. J.* **43,** 314–318.

Malinowski E. R. (1977) Determination of the number of factors and the experimental error in a data matrix. *Anal. Chem.* **49,** 612–617.

Malinowski E. R. and McCue M. (1977) Qualitative and quantitative determination of suspected components in mixtures by target transformation factor analysis of their mass spectra. *Anal. Chem.* **49,** 284–287.

Miesch A. T. (1976) *Q*-mode factor analysis of geochemical and petrologic data matrices with constant row sums. U.S. Geological Survey Professional Paper 574-G, U.S. Government Printing Office, Washington, DC.

Miller M. S., Friedlander S. K. and Hidy G. M. (1972) A chemical element balance for the Pasadena aerosol. *J. Colloid Interface Sci.* **39,** 165–176.

Mroz E. J. (1976) The Study of the Elemental Composition of Particulate Emissions from an Oil-Fired Power Plant. Ph.D. Thesis, University of Maryland, College Park, MD.

Mosteller F. (1971) The jacknife. *Rev. Int. Stat. Inst.* **39,** 363–368.

Mosteller F. and Tukey J. (1977) *Data Analysis and Regression*. Addison-Wesley Publishing Co., Reading, MA.

Ondov J. M. (1974) A Study of Trace Elements on Particulates from Motor Vehicles. Ph.D. Thesis, University of Maryland, College Park, MD.

Prinz B. and Stratmann H. (1968) The possible use of factor analysis in investigating air quality. *Staub-Reinhalt Luft.* **28,** 33–39.

Roscoe B. A. and Hopke P. K. (1981a) Comparison of weighted and unweighted target transformation rotations in factor analysis. *Comput. Chem.* **5,** 1–7.

Roscoe B. A. and Hopke P. K. (1981b) Error estimation of factor loadings and scores obtained with target transformation factor analysis. *Anal. Chim. Acta* **132,** 89–97.

Roscoe B. A., Cheng C. Y. and Hopke P. K. (1984) Comparison of the target transformation factor analysis of coal composition data with x-ray diffraction data. *Anal. Chim. Acta* **160,** 121–134.

Severin K. G., Roscoe B. A. and Hopke P. K. (1983) The use of factor analysis in source determination of particulate emissions. *Particulate Sci. Technol.* **1,** 189–196.

Thomae S. C. (1977) Size and Composition of Atmospheric Particles in Rural Areas Near Washington, D.C. Ph.D. Thesis, University Of Maryland, College Park, MD.

Thurston G. D. and Spengler J. D. (1985) A quantitative assessment of source contributions to inhalable particulate pollution in metropolitan Boston. *Atmospheric Environment* **19,** 9–25.

Watson J. G., Jr (1979) Chemical Element Balance Receptor Methodology for the Sources of Fine and Total Suspended Particulate Matter in Portland, Oregon. Ph.D. Thesis, Oregon Graduate Center, Beaverton, OR.

Watson J. G., Cooper J. A. and Huntzicker J. H. (1984) The effective variance weighting for least squares calculations applied to the mass balance receptor model. *Atmospheric Environment* **18,** 1347–1355.

Weiner P. H., Malinowski E. R. and Levinston A. R. (1970) Factor analysis of solvent shifts in proton magnetic resonance. *J. phys. Chem.* **74,** 4537–4542.

Zhou D., Chang T. and Davis J. C. (1983) Dual extraction of *R*-mode and *Q*-mode factor solutions. *Int. J. Math. Geol. Assoc.* **15,** 581–606.

NOTICE

This document has been reviewed in accordance with U.S. Environmental Protection Agency policy and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

This article documents research that was conducted under the auspices of the U.S. Environmental Protection Agency (USEPA). As an arm of the government, the USEPA has a paid up, non-exclusive, irrevocable, worldwide license to reproduce, prepare derivative works, and distribute to the public all of the material contained herein, and to grant others acting on its behalf the right to do so. The USEPA grants to the National Technical Information Service, Department of Commerce, the right to act as its agent in the distribution of this material.